DOCUMENT RESUME

ED 030 949 EA 002 271

By-Tyler, Ralph W.

Changing Concepts of Educational Evaluation.

Pub Date 67

Note-11p.; Pages 13-18 in PERSPECTIVES OF CURRICULUM EVALUATION, AERA Monograph Series on Curriculum Evaluation, edited by Ralph W. Tyler, And Others, Rand McNally & Co., Chicago, 1967.

Available from-Rand McNally & Company, Box 7600, Chicago, Illinois 60680 (Complete document 102 pages, $2.00).

EDRS Price MF-$0.25 HC-$0.65

Descriptors-Diagnostic Tests, *Educational Programs, *Educational Research, *Evaluation Methods, *Evaluation Techniques, Factor Analysis, Learning Processes, Measurement

The acceleration of educational research has resulted in an array of concepts, research instruments, and methods which demand clarification and integration, because new conditions and assumptions have been introduced without considering their effect upon the educational process. Special attention needs to be given to the development of evaluation procedures for such objects as the educational progress of large populations, educational innovations and the utility of the instruments by which they are evaluated, the use of various forms of factor analysis, the educability of humans generally, the whole area of diagnostic testing, the role of the learner, and the nature of knowledge. If educational evaluation is to make a positive contribution, it must be in harmony with the basic assumptions of educational programs already in operation. (JK)

AERA MONOGRAPH SERIES
ON CURRICULUM EVALUATION

ED 030949

1

PERSPECTIVES OF CURRICULUM EVALUATION

Ralph Tyler
Robert Gagne
Michael Scriven

EA 002 271

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION
MONOGRAPH SERIES
ON
CURRICULUM EVALUATION

Committee on Curriculum Evaluation:
Harold Berlak
Leonard S. Cahen
Richard A. Dershimer
Christine McGuire
Jack C. Merwin
Ernst Z. Rothkopf
James P. Shaver
Robert E. Stake

Coordinating Editor
Robert E. Stake

Editorial Associates:
J. Myron Atkin
Peter A. Taylor

Editorial Consultants for this Issue:
John B. Gilpin
J. Thomas Hastings
Thomas O. Maguire

American Educational Research Association
1126 Sixteenth Street, N. W.
Washington, D. C. 20036

Perspectives of Curriculum Evaluation

Ralph W. Tyler

Center for Advanced Study in the Behavioral Sciences

Robert M. Gagné

University of California, Berkeley

Michael Scriven

Indiana University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

$1 2.00

Rand McNally & Company
Chicago

Box 7600
60680

RAND McNALLY EDUCATION SERIES

B. OTHANEL SMITH, *Series Editor*

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Barbara Siefken, Rand McNally & Company TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

Copyright © 1967 by Rand McNally & Company
All rights reserved
Printed in U.S.A. by Rand McNally & Company
Library of Congress Catalog Card Number 66-30794

Contents

	Page
Toward a Technology for the Evaluation of Educational Programs: Robert E. Stake	1
Changing Concepts of Educational Evaluation: Ralph W. Tyler	13
Curriculum Research and the Promotion of Learning: Robert M. Gagné	19
The Methodology of Evaluation: Michael Scriven	39
Aspects of Curriculum Evaluation: A Synopsis: J. Stanley Ahmann	84
Bibliography	90

ED030949

Changing Concepts of Educational Evaluation

Ralph W. Tyler[1]

Center for Advanced Study in the Behavioral Sciences, Stanford, California

I have chosen this topic because it seems to me to deal with a problem likely to be faced in many areas of educational research as larger support enables us to move more rapidly and more comprehensively in developing scientific knowledge about education. My thesis is: The accelerating development of research in the area of educational evaluation has created a collection of concepts, facts, generalizations, and research instruments and methods that represent many inconsistencies and contradictions because new problems, new conditions, and new assumptions are introduced without reviewing the changes they create in the relevance and logic of the older structure.

To illustrate this, I should like to cite first our experience in the project concerned with assessing the progress of education. The purpose is to appraise the educational progress of large populations in order to provide the public with dependable information to help in the understanding of educational problems and needs and to guide in efforts to develop sound public policy regarding education. This type of evaluation is not focused upon individual students, classrooms, schools, or school systems, but is to furnish over-all information about the educational attainments of large numbers of people. Although the purpose is not identical with that of current achievement testing programs, we thought that available tests and/or test items would serve our purposes, but this turned out not to be the case.

Because current achievement tests seek to measure individual differences among pupils taking the tests, the items are concentrated on those which differentiate among the children. Exercises which all or nearly all can do, as well as those which only a

EA 002 271

[1]A version of this paper was given as an invited address at the American Educational Research Association Annual Meeting, Chicago, February 17, 1966.

very few can do, are eliminated because these do not give much discrimination. But, for the purposes of assessing the progress of education, we need to know what all, or almost all, of the children are learning and what the most advanced are learning as well as what is being learned by the middle or "average" children. To obtain exercises of this sort is a new venture for most test constructors.

Because of the prevailing concept of measuring achievement in terms of the relative performance of individuals within a group, the difficulty level of test items is often manipulated by the item writer who changes the wording of the stem of the exercise, or of some of the multiple answers, without seeming to realize that in many cases this changes the nature of the behavior being appraised. Test practice and, to some extent, theory have been based on assumptions that are acceptable only for certain kinds of work.

We also find that little theory has been formulated or techniques devised to aid in the construction of relatively homogeneous samples of exercises faithfully reflecting an educational objective. The typical reliability coefficients refer to individual scores and not to the homogeneity of a given level of behavior. Our project will not compute individual scores, but the following sorts of things will be reported:

> For the sample of seventeen-year old boys of higher socioeconomic status from rural and small town areas of the Midwest region, it was found that:
>
> 93% could read a typical newspaper paragraph like the following.
> 76% could write an acceptable letter ordering several items from a store like the following.
> 52% took a responsible part in working with other youth in playground and community activities like the following.
> 24% had occupational skills required for initial employment.

Evaluation exercises on which such reports can be made must represent an acceptable degree of homogeneity so that, for example, the "newspaper paragraph" referred to above is typical of dozens of other paragraphs that are shown to represent the same difficulty level. We found that some people were interpreting current test items as though they were reliable samples of a given level of difficulty. Thus, reporters would comment on the fact that on a certain test only 12% of the students knew that Boise was the capitol of Idaho. If only 12% of those tested got this exercise right, this performance is likely to be idiosyncratic since neither do

current tests seek to establish representative and reliable samples of what is known by only 12% of students nor do they comprehensively sample knowledge of state capitols. Our present instruments are products of assumptions and conditions that do not properly apply to some of our current needs for evaluation.

Another series of concepts, procedures, and instruments needs reexamination because of the current emphasis upon innovation at all levels of education. We can no longer depend so heavily upon the assumption that success in schools or colleges as they are now operating is an acceptable criterion for validating a measuring instrument. The task of the elementary school is now recognized as that of reaching all children, including the 15 to 20 per cent who have not been making appreciable progress in learning before. Our society can find constructive places for no more than 5 to 10 per cent of its people who are unskilled and untutored. The task of the high school is now recognized as that of educating a very large proportion of youth, including the 25 to 35 per cent who have not been making substantial progress in earlier years. The changing structure of the labor force, the higher requirements for intelligent citizenship both make this demand. Finally, the task of the college and university is to reach at least 50 per cent of our youth in order that our complex, industrial society can continue to develop.

These demands make untenable the assumption that there is a large pool of humanity from which the cream is to be skimmed off for certain educational or occupational purposes, and that failures in educational institutions are principally due to poor selection of students. On the other side of the coin, we now see that schools and colleges, like other institutions, become program-centered, losing their orientation toward their clients. Most institutions begin as responses to the need of certain clients for services. As years go by, programs are developed that are reasonably acceptable to the clients they have been serving. Then the institution is likely to believe that its program is its raison d'être rather than the need for its services. When this program-worship stage is reached, the institution seeks to find clients who like the program and can get along with it, and to deny admission to others. After a time, the terminology develops that those not admitted are "poor students," "not intelligent," not of "college calibre." In many cases, as in the founding of the Land-Grant Colleges, new institutions have to be established to serve the clients rejected by the older ones.

The current climate in this country is to seek innovation, to get the institutions active in learning how to serve their new clients. Evaluative instruments for this purpose must avoid using criteria based upon the current judgments of schools and colleges because

this criterion perpetuates the conviction that these institutions are, at present, satisfactory for the tasks to be done.

In place of the older criteria and the dependent procedures we need new concepts of educational readiness, strengths on which to build, deficiencies to be attacked, and the like. These new concepts must be based on the assumption of dynamic potential in all or almost all human beings. The evaluation task is to describe or measure phases of this potential and difficulties to be surmounted that can help the individual and the educational institution in improving student learning.

This raises still another area for reexamination, the uncritical use of various forms of factor analysis. When Truman Kelley wrote his monumental volume, *Crossroads in the Mind of Man* (1928), nearly 40 years ago, scientific studies of human heredity were very limited. Hence, it was not a result of gross anti-intellectualism or ignorance that many psychologists and educators thought of factors as something inherent in the neural mechanism. However, this type of naïvete still continues in spite of the great advances that are being made in the sciences of human genetics and neurology. Because many of us think of "factors" obtained from tests as indicators of basic neural connections rather than as learned similarities or generalizations, we tend to use factor scores or to restrict the range of types of evaluation exercises in cases where they are not appropriate or may be misleading, as well as in cases where their employment increases the efficiency of measurement.

Related to this is the changing conception of the plasticity and educability of human organisms generally. The initial work on transfer of training has left a legacy in which both teaching and testing are frequently based on a stochastic concept of learning. The earlier work of Judd (1921), Freeman (1917), and their students on generalization and teaching for transfer was largely overlooked until the emphasis given more recently by Bruner (1960) and others on the structure of the disciplines and the organized nature of cognitive learning. In both theory and practice, we who work on educational evaluation are still guided by procedures and instruments that treat items as units and learning measures largely as the sum of specific, unorganized bits.

The whole area of diagnostic testing has largely been neglected in practice although much of the basic theory was outlined in the milestone volume *Educational Measurement*, edited by E. F. Lindquist and published by the American Council of Education (Lindquist, 1951). In its Chapter 1, pp. 37-38, the late Walter Cook listed the following general criteria for diagnostic tests:

> (1) They must be an integral part of the curriculum, emphasizing and clarifying the important objectives. (2) The test items should require responses to be made to situations approximating as closely as possible the functional. (3) The tests must be analytical and based on experimental evidence of learning difficulties and misunderstandings. (4) The tests should reveal the mental processes of the learner sufficiently to detect points of error. (5) The tests should suggest or provide specific remedial procedures for each error detected. (6) The tests should be designed to cover a long sequence of learning systematically. (7) The tests should be designed to check forgetting by constant review of difficult elements, as well as to detect faulty learning. (8) Pupil progress should be revealed in objective terms.

In Chapter 9 of that volume, pp. 266-67, Frederick Davis, in discussing item selection, comments:

> The difficulty of individual items is not an important consideration when items are selected for mastery tests. . . . Mastery tests are not intended to provide scores that will rank students in terms of their knowledge or ability; rather they are designed to separate students into two groups, those who know certain basic facts, principles or operations, and those that do not know them. For this reason, the items in a mastery test are so chosen that nearly every pupil who has reached a predetermined level of achievement can answer them all correctly.

In spite of these published statements, there are very few tests available meeting these conditions for diagnostic purposes. Now that high-speed computers and electronic data processing make individual diagnosis, recording, and treatment feasible, teachers do not have appropriate evaluation instruments to guide greater individualization of instruction. We are still so obsessed with the ranking of individuals on the basis of scores that we have not developed adequately the tools and procedures required. Theory and practice need to be reexamined in terms of present conditions and opportunities.

Perhaps the most basic assumptions we make in educational evaluation are those that deal with our concept of the role of the learner in learning and related notions about the nature of knowledge. If we conceive of the learner as one who is learning to make appropriate responses to situations outside his control, we are likely to think of learning as a kind of conditioning in which the only choice open to the learner is to react "correctly" or to refuse to respond. On the other hand, learning may be viewed as a process

by which the learner develops a behavior that enables him to deal satisfactorily with the situation which he confronts in a way that more nearly achieves his purposes. The cartoon showing one rat telling another that he has got the psychologist under control because the psychologist gives him food whenever he presses a lever, seems humorous to us, but it illustrates the possibility of the learner devising ways to manipulate the situations he encounters rather than views him as one who must respond as the text or test requires or be penalized.

John Dewey was speaking of this when he characterized a good learning situation as one which requires the learner to make certain adaptations to those conditions beyond his control and to modify other conditions so as to serve his ends. Dewey considered a situation in which there was no freedom for the learner to reconstruct conditions as uneducative, as was also one in which he could manipulate all conditions according to his whim or fancy.

Related to this is the unstated view about knowledge. Is knowledge something "out there" which we must find out, remember, and follow, or is knowledge a product of human efforts to make sense out of the world, to accomplish certain tasks and to enhance human satisfactions? Knowledge as a continuing human effort is something which the learner must help to construct.

These different conceptions make a difference in the way achievement test exercises are designed, the directions written, and the "response-set" stimulated. We have not recently reviewed our theory and practice in educational evaluation to assure ourselves that they are in harmony with the basic assumptions on which current educational programs are operating.

The illustrations that I have been presenting are not exhaustive, but I hope that they have provided some indication of the meaning of the thesis with which I began—"The accelerating development of research in the area of educational evaluation has created a collection of concepts, facts, generalizations, and research instruments and methods that represent many inconsistencies and contradictions because new problems, new conditions, and new assumptions are introduced without reviewing the changes they create in the relevance and logic of the older structure." Before this mixed vegetation becomes a jungle, can we not establish the equivalent of the Physical Science Study Committee to sort out and arrange our materials in a more ordered fashion, eliminating obsolete notions, clarifying and strengthening the framework of our field?